1 Introduction

Let us start by recalling the online gradient descent for optimizing convex functions. Remember the set up: given a fixed $\epsilon > 0$, we present at each time step $t$ a vector $x_t$ in a closed convex set $K \subseteq \mathbb{R}^n$, the adversary will then choose a function $f_t : K \rightarrow \mathbb{R}$ which is convex and smooth. We also assume $f_t$ is $G$-Lipschitz with respect to $\| \cdot \|_2$, which means

$$\frac{f_t(x) - f_t(y)}{\| x - y \|_2} \leq G$$

for all distinct $x, y \in K$, or equivalently $\| \nabla f_t(x) \|_2 \leq G$ for all $x \in K$.

We showed that for any $x^* \in K$, a slightly modified variant of the gradient descent algorithm, starting from a point $x_0 \in K$ with $\| x_0 - x^* \|_2 \leq D$ and after $T$ steps, produces $x_1, \ldots, x_T$ such that

$$x_i \in K \quad \text{for} \quad i = 1, \ldots, T,$$

and

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x^*) + \frac{\eta \sum_{t=1}^{T} \| \nabla f_t(x_t) \|_2^2}{2} + \frac{\| x^* - x_0 \|_2^2}{2\eta}. \quad (15.1)$$

Set $\eta = \frac{D}{G\sqrt{T}}$ to get

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x^*) + \frac{GD}{\sqrt{T}}. \quad (15.2)$$

Then, we can set $T = \left( \frac{GD}{\epsilon} \right)^2$ and $\hat{x} = \frac{1}{T} \sum_{i=1}^{T} x_i$ to get

$$\sum_{t=1}^{T} f_t(\hat{x}) \leq \sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x^*) + \frac{\epsilon}{\text{regret}}. \quad (By \ convexity \ of \ f_t)$$

$$\leq \sum_{t=1}^{T} f_t(x^*) + \frac{\epsilon}{\text{regret}} \quad (By \ 15.2)$$

Notice that this gradient descent algorithm works for all convex functions over convex bodies, as for Multiplicative Weight (MW) algorithm which only works for linear functions and over $\Delta_n = \{ x \in \mathbb{R}_+^n : \sum_{i=1}^{n} x_i = 1 \}$, i.e. the simplex in $\mathbb{R}^n$. Let us illustrate this difference in more details in the following example to motivate the topic for today’s lecture.

**Example 15.1.** Suppose $f_t : \Delta_n \rightarrow \mathbb{R}$ and $f_t(x) = \langle \ell_t, x \rangle$, where $\ell_t \in [-1,1]^n$ for $t = 1, \ldots, T$. Notice that for all $t = 1, \ldots, T$, function $f_t$ is $\sqrt{n}$-Lipschitz, and for any $x_0 \in \Delta_n$ we have $\| x_0 - x^* \|_2 \leq \sqrt{2}$ for all $x^* \in \Delta_n$. Hence, applying the online gradient descent method for $T = \left( \frac{\sqrt{2} \sqrt{n}}{\epsilon} \right)^2 = \frac{2n}{\epsilon^2}$ outputs a solution $\hat{x}$ with regret at most $\epsilon$.

On the other hand, this problem is an MW problem. Hence, we can apply Hedge algorithm for $T = \frac{\ln n}{\epsilon}$ steps to get a regret of at most $\epsilon$.

Therefore, gradient descent needs significantly more steps to be able to guarantee an $\epsilon$ regret compared to Hedge algorithm.
2 Norms and their Duals

In the previous section we described a gradient descent method which relied on the Euclidean norm \( \| \cdot \|_2 \). Today we will try different norm functions to see if we can overcome the shortcoming of gradient descent that was mentioned in Example 15.1. First we need to formally define a norm and its dual.

**Definition 15.2.** A function \( \| \cdot \| : \mathbb{R}^n \to \mathbb{R} \) is a norm if

1. If \( \| x \| = 0 \) for \( x \in \mathbb{R}^n \), then \( x = 0 \);
2. for \( \alpha \in \mathbb{R} \) and \( x \in \mathbb{R}^n \) we have \( \| \alpha x \| = |\alpha| \| x \| \); and
3. for \( x,y \in \mathbb{R}^n \) we have \( \| x + y \| \leq \| x \| + \| y \| \).

**Example 15.3.** \( \ell_p \)-norm for \( p \in \mathbb{Z}_+ \) is defined as \( \| x \|_p = \left( \sum_{i=1}^n x_i^p \right)^{\frac{1}{p}} \) for \( x \in \mathbb{R}^n \). Also \( \ell_\infty \)-norm is defined as \( \| x \|_\infty = \max_{i=1,...,n} x_i \) for \( x \in \mathbb{R}^n \). See Figure 15.1 for further illustration.

![Figure 15.1: The unit ball in \( \ell_1 \)-norm (Green), \( \ell_2 \)-norm (Blue), and \( \ell_\infty \)-norm (Red).](image)

**Definition 15.4.** Let \( \| \cdot \| \) be a norm. Then the dual norm of \( \| \cdot \| \) is a function \( \| \cdot \|_* \) defined as

\[
\| y \|_* = \sup \{ \langle x, y \rangle : \| x \| \leq 1 \}.
\]

**Corollary 15.5.** For \( x,y \in \mathbb{R}^n \), we have \( \langle x, y \rangle \leq \| x \| \| y \|_* \).

*Proof.* Assume \( \| x \| \neq 0 \), otherwise both sides are 0. Since \( \| x \|_{\| x \|} = 1 \), we have \( \langle \| x \|_{\| x \|}, y \rangle \leq \| y \|_* \). \( \square \)

**Example 15.6.** The dual norm of \( \ell_2 \)-norm is \( \ell_2 \)-norm. The dual norm of \( \ell_1 \)-norm is the \( \ell_\infty \)-norm.

**Theorem 15.7.** The dual norm of \( \ell_p \)-norm \( \| \cdot \|_p \) is \( \ell_q \)-norm \( \| \cdot \|_q \), where \( \frac{1}{p} + \frac{1}{q} = 1 \).

**Theorem 15.8.** We have \( (\| \cdot \|_*)_* = \| \cdot \| \), for \( \| \cdot \| \) defined on a finite dimension space.

3 Online Mirror Descent

We now review the mirror descent algorithm introduced by Nemirovski and Yudin [NY78]. Recall in gradient descent method in each step we set \( x_{t+1} = x_t - \eta \nabla f_t(x_t) \). Note that \( \nabla f_t \) is a function in the dual space. We often overlook this fact since in the gradient descent method we work in \( \mathbb{R}^n \) with \( \ell_2 \)-norm, and this normed space is in fact self-dual. However, Example 15.1 suggests that \( \ell_2 \)-norm might not be the “right” norm. To this end, we define a refined version of lipschitz continuity for a norm \( \| \cdot \| \).
Definition 15.9. Let $f$ be a differentiable function. Then $f$ is $G$-Lipschitz with respect to $\| \cdot \|$ if
\[ \| \nabla f(x) \|_* \leq G \text{ for all } x \in \mathbb{R}^n. \]

Since $\nabla f_t$ is a function in the dual space $-\eta \nabla f_t(x_t)$ is a step in the dual space. Hence, we need to map our current point $x_t$ to a point in the primal space, namely $\theta_{t+1}$. After taking the gradient step, $\theta_{t+1} = \theta_t - \eta \nabla f_t(x_t)$ we still have to map $\theta_{t+1}$ back to a point in the primal space $x'_{t+1}$. Similar to gradient descent $x'_{t+1}$ might not be in the closed convex feasible region $K$, so we need to project $x'_{t+1}$ back to a “close” $x_{t+1}$ in $K$. This was an informal description of the mirror descent algorithm (See Figure 15.2).

![Figure 15.2: The four basic steps in each iteration of the mirror descent algorithm](image-url)

To justify the appellation of the algorithm, notice that the dual space acts as a mirror to the primal space. That is why we call the functions that map $x_t$ to $\theta_t$ and $\theta_{t+1}$ to $x'_{t+1}$ the mirror maps. To find a suitable mirror map, we need to define $\alpha$-strongly convex function with respect to a norm $\| \cdot \|$.

Definition 15.10. Convex and differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ is $\alpha$-strongly convex with respect to $\| \cdot \|$ if
\[ h(y) \geq h(x) + \langle \nabla h(x), y - x \rangle + \frac{\alpha}{2} \| y - x \|^2. \]

Example 15.11. Function $h_1 : \mathbb{R}^n \to \mathbb{R}$ defined as $h_1(x) = \frac{1}{2} \| x \|^2_2$ is 1-strongly convex with respect to $\| \cdot \|_2$.

Example 15.12. Function $h_2 : \mathbb{R}^n \to \mathbb{R}$ defined as $h_2(x) = \sum_{i=1}^n x_i \log x_i$ is $\frac{1}{\ln 2}$-strongly convex with respect to $\| \cdot \|_1$. Function $h_2$ is the negative entropy function.

Let $h : \mathbb{R}^n \to \mathbb{R}$ be an $\alpha$-strongly-convex function wrt $\| \cdot \|$. Then, we will use $\nabla(h) : \mathbb{R}^n \to \mathbb{R}^n$ as our mirror map. Thus, we will set $\theta_t = \nabla h(x_t)$, and $x'_{t+1} = (\nabla h)^{-1}(\theta_{t+1})$. See Figure 15.2.

Example 15.13. Recall function $h_1$ is Example 15.11. We have $\nabla h_1(x) = x$, and $(\nabla h_1)^{-1}(\theta) = \theta$.

Example 15.13 gives a nice intuition why the gradient descent algorithm works within the primal and dual space unnoticed.

Example 15.14. Consider function $h_2$ in Example 15.12. We have $\nabla h_2(x)_i = (\ln x_i + 1)_i$, and $(\nabla h_2)^{-1}(\theta)_i = (e^{\theta_i-1})_i$. 

3
As mentioned before, the mirror descent algorithm is basically similar to gradient descent when we are working in $\mathbb{R}^n$ normed with $\|\cdot\|_2$, and when the mirror map is $\nabla h_1$. Hence, we will explain the algorithm when we are on $\mathbb{R}^n$ normed with $\|\cdot\|_1$ and mirror map $\nabla h_2$. For simplicity, we refer to $x_t, x_{t+1}, \theta_t,$ and $\theta_{t+1}$ by $x, x', x^+, \theta,$ and $\theta^+$, respectively.

(i) Start with $x$ and compute $\theta = (\ln x_i + 1)_i$, i.e. map $x$ to $\theta$ using the mirror map $\nabla h_2$ to the dual space.

(ii) Set $\theta^+ = (\theta - \eta \nabla f_t(x)) = (\ln x_i + 1 - \eta \nabla f_t(x)_i)_i$, i.e. take the gradient step in the dual space.

(iii) Find $x' = (e^{\ln x_i - \eta (\nabla f_t(x)_i)}_i = (x_i \cdot e^{-\eta (\nabla f_t(x)_i)})_i$, i.e. map $\theta^+$ back to the primal space.

Remember Example 15.1 where $f_t(x) = \langle \ell_t, x \rangle$, in this case $\nabla f_t = \ell_t$, so the mirror descent algorithm finds $x' = (x_i e^{-\eta \ell_t}_i)_i$, which is similar to Hedge algorithm.

There is still one missing step in the algorithm:

(iv) Project $x'$ back to point $x^+$ in the feasible region $K$.

In order to do this, we need to define Bregman distance.

**Definition 15.15.** The Bregman distance of $x$ and $y$ with respect to function $h$, denoted by $D_h(y\|x)$ is

$$h(y) - h(x) - \langle \nabla h(x), y - x \rangle.$$

Figure 15.3 describes the Bregman distance geometrically for $h : \mathbb{R} \to \mathbb{R}$.

![Figure 15.3: $D_h(y\|x)$ for function $h : \mathbb{R} \to \mathbb{R}$](image)

We can now define the notation of Bregman projection.

**Definition 15.16.** The Bregman projection of point $x'$ onto convex set $K$ is

$$x^+ = \arg \min_{x \in K} D_h(x\|x').$$

**Example 15.17.** Consider function $h_1(x) = \frac{1}{2} \|x\|_2^2$ from Example 15.11. Then

$$D_{h_1}(y\|x) = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|x\|_2^2 - \langle x, y - x \rangle$$

$$= \frac{1}{2} \|y\|_2^2 + \frac{1}{2} \|x\|_2^2 - \langle x, y \rangle$$

$$= \frac{1}{2} \|y - x\|_2^2.$$
Therefore, when we apply the mirror descent algorithm with $\ell_2$-norm and mirror function $h_1$, the projection step is exactly similar to the projection step in gradient descent. This is because for $h_1$, Bregman distance basically similar to the Euclidean distance.

**Example 15.18.** For function $h_2(x) = \sum_{i=1}^{n} x_i \ln x_i$ from Example 15.12, we have

$$D_{h_2}(y\|x) = \sum_{i=1}^{n} y_i \ln y_i - \sum_{i=1}^{n} x_i \ln x_i - \sum_{i=1}^{n} (\ln x_i + 1)(y_i - x_i)$$

$$= -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \ln \frac{y_i}{x_i},$$

where $KL(y\|x)$ is known as the Kullback-Leibler divergence. Now in the case of $\ell_1$-norm with mirror map $h_2$, step (iv) is

$$(iv) \quad x^+ = \left( \frac{x_i e^{\eta l_i}}{\sum_{j=1}^{n} x_j e^{\eta l_j}} \right)_i,$$

i.e. take Bregman projection of $x'$ onto the feasible region (the unit simplex $\Delta_n$) with respect to Bregman distance $D_{h_2}$.

### 4 Analysis

We prove the following theorem.

**Theorem 15.19.** Let $f_1, \ldots , f_T$ be convex and differentiable functions, $\| \cdot \|$ be a norm function, and $h$ be an $\alpha$-strongly convex function with respect to $\| \cdot \|$, then the mirror descent algorithm starting with $x_0$ and taking constant step size $\eta$ in every iteration, produces $x_1, \ldots , x_T$ such that

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{n} f_t(x^*) + \frac{D_h(x^\|x_0)}{\eta} + \frac{\eta \sum_{t=1}^{T} \|\nabla f_t(x_t)\|^2}{2\alpha},$$

for all $x^*$ \hspace{1cm} (15.3)

Before proving Theorem 15.19, let us take a look at Inequality 15.3 in the two cases we discussed at length in the previous section.

If $\| \cdot \|$ is $\ell_2$-norm and $h$ is function $h_1$ from Example 15.11, then Inequality 15.3 becomes

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{n} f_t(x^*) + \frac{\|x^* - x_0\|^2}{2\eta} + \frac{\eta \sum_{t=1}^{T} \|\nabla f_t(x_t)\|^2}{2},$$

for all $x^*$,

which is Inequality 15.1.

If $\| \cdot \|$ is $\ell_1$-norm and $h$ is function $h_2$ from Example 15.12, then Inequality 15.3 becomes

$$\sum_{t=1}^{T} \langle \ell_t, x_t \rangle \leq \sum_{t=1}^{n} \langle \ell_t, x^* \rangle + \frac{\sum_{i=1}^{n} x_i^* \ln \frac{x_i^*}{x_i}}{2\eta} + \frac{\eta \sum_{t=1}^{T} \|\ell_t\|_\infty^2}{2},$$

for all $x^* \in \Delta_n$.

Since $\|\ell_t\|_\infty \leq 1$, we have

$$\sum_{t=1}^{T} \langle \ell_t, x_t \rangle \leq \sum_{t=1}^{T} \langle \ell_t, x^* \rangle + \frac{\ln n}{2\eta} + \frac{\eta T}{2},$$

for all $x^* \in \Delta_n$. 

5
Proof of Theorem 15.19. Define potential \( \Phi_t = \frac{D_h(x^* \| x_t)}{\eta} \). The amortized cost at time \( t \) is
\[
f_t(x_t) - f_t(x^*) + (\Phi_{t+1} - \Phi_t).
\]
Now
\[
\Phi_{t+1} - \Phi_t = \frac{1}{\eta} \left( D_h(x^* \| x_{t+1}) - D_h(x^* \| x_t) \right)
\]
\[
= \frac{1}{\eta} \left( h(x_t) - h(x_{t+1}) - (\nabla h(x_{t+1}), x^* - x_{t+1}) - h(x^*) + h(x_t) + (\nabla h(x_t), x^* - x_t) \right)
\]
\[
= \frac{1}{\eta} \left( h(x_t) - h(x_{t+1}) - (\theta_t - \eta \nabla f_t(x_t), x^* - x_{t+1}) + (\theta_t, x^* - x_t) \right)
\]
\[
= \frac{1}{\eta} \left( h(x_t) - h(x_{t+1}) - (\theta_t, x_t - x_{t+1}) + \eta(\nabla_t, x^* - x_{t+1}) \right)
\]
\[
\leq \frac{1}{\eta} \left( \frac{\alpha}{2} \|x_{t+1} - x_t\|^2 + \eta(\nabla_t, x^* - x_{t+1}) \right) \quad \text{(By \( \alpha \)-strong convexity of} \ h \ \text{wrt to} \ \| \cdot \|)
\]
Plug this back to 15.4
\[
f_t(x_t) - f_t(x^*) + (\Phi_{t+1} - \Phi_t) \leq f_t(x_t) - f_t(x^*) + \frac{\alpha}{2\eta} \|x_{t+1} - x_t\|^2 + (\nabla_t, x^* - x_{t+1})
\]
\[
\leq f_t(x_t) - f_t(x^*) + \frac{\alpha}{2\eta} \|x_{t+1} - x_t\|^2 + (\nabla_t, x_t - x_{t+1})
\]
\[
\leq \frac{\alpha}{2\eta} \|x_{t+1} - x_t\|^2 + \|\nabla_t\|_* \|x_t - x_{t+1}\| \quad \text{(By Corollary 15.5)}
\]
\[
\leq \frac{\alpha}{2\eta} \|x_{t+1} - x_t\|^2 + \frac{1}{2} \left( \frac{\eta}{\alpha} \|\nabla_t\|_*^2 + \frac{\alpha}{\eta} \|x_t - x_{t+1}\|^2 \right) \quad \text{(By AM-GM)}
\]
\[
\leq \frac{\eta}{2\alpha} \|\nabla_t\|_*^2.
\]
Thus,
\[
\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \leq \Phi_0 - \Phi_{T+1} + \sum_{t=1}^{T} \frac{\eta}{2\alpha} \|\nabla_t\|_*^2
\]
\[
\leq \Phi_0 + \sum_{t=1}^{T} \frac{\eta}{2\alpha} \|\nabla_t\|_*^2
\]
\[
\leq \frac{D_h(x^* \| x_0)}{\eta} + \frac{\eta}{2\alpha} \sum_{t=1}^{T} \|\nabla_t\|_*^2.
\]

5 Mirror Descent as Prox version of Gradient Descent

In this lecture, we reviewed mirror descent algorithm as a gradient descent scheme where we do the gradient step in the dual space. A shorter (but less intuitive) description of mirror descent in the following.
Algorithm 1 Mirror Descent Algorithm

\begin{algorithm}
\textbf{for} $t \leftarrow 0$ to $T - 1$ \textbf{do}
\begin{align*}
    x_{t+1} &\leftarrow \arg \min_{x \in K} \{ \eta \langle \nabla f_t(x_t), x \rangle + D_h(x \| x_t) \}
\end{align*}
\end{algorithm}

References
